Fall 2023 Econometrics I: Lecture 1

Review – Matrix Algebra, Probability & Statistics

Eric S. Lin

Department of Economics, National Tsing Hua University

September 10, 2024

Purposes

- Important tools
  - simple to multiple regression model using succinct expressions
  - for complicated models
- Important concepts
  - probability (likelihood function, conditional prob)
  - statistics (testing procedure, size)
- Different from other disciplines
  - linear algebra
  - don't go too far
- Dirty your hand
  - listening is not enough
  - getting pieces of paper

Reference

- *Econometric Analysis*, 8th Edition, William H. Greene, Stern School of Business, New York University, 2021, Pearson
- PART VI. Appendices (online)
  - Appendix A: Matrix Algebra
  - Appendix B: Probability and Distribution Theory

Basics

- Vector, Matrix, Matrix addition, and Matrix subtraction
  - special case of a matrix: column vector, row vector
- Matrix multiplication
  If A is of order $m \times n$ [(rows) by (columns), size, dimension] and B is of order $n \times p$, then $C = AB$ is of order $m \times p$. We say that matrices A and B are conformable.
  - use a_{ij} to denote a particular element of matrix A
- Matrix partition

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Matrix of Macro Data

	Column				
Row	1 Year	2 Consumption (billions of dollars)	3 GNP (billions of dollars)	4 GNP Deflator	5 Discount Rate (N.Y Fed., avg.)
1	1972	737.1	1185.9	1.0000	4.50
2	1973	812.0	1326.4	1.0575	6.44
3	1974	808.1	1434.2	1.1508	7.83
4	1975	976.4	1549.2	1.2579	6.25
5	1976	1084.3	1718.0	1.3234	5.50
6	1977	1204.4	1918.3	1.4005	5.46
7	1978	1346.5	2163.9	1.5042	7.46
8	1979	1507.2	2417.8	1.6342	10.28
9	1980	1667.2	2633.1	1.7864	11.77

Source: Data from the *Economic Report of the President* (Washington, D.C.: U.S. Government Printing Office, 1983).

Partition

Example

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11(2\times 3)} & A_{12(2\times 1)} \\ A_{21(1\times 3)} & A_{22(1\times 1)} \end{bmatrix}$$

$$A_{11(2\times 3)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

$$A_{21(1\times 3)} = \begin{bmatrix} a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Matrix Inverse and Transpose

- Matrix inverse

$$A^{-1} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$$

A matrix which has an inverse is called to be nonsingular

  - invertible matrix
  - property: $(AB)^{-1} = B^{-1}A^{-1}$

- Transpose Operator
  - turns columns into rows (and vice versa)
  - suppose x is a column vector
  - $x' = x^T = [x_1, x_2, ..., x_n]$: row vector
  - $x'x = \sum_{i=1}^{n} x_i^2$: a scalar
  - $A' = (a_{ji})_{n \times m}$ [recall that $A = (a_{ij})_{m \times n}$]

Special Matrices

- Square matrix
  - an $m \times n$ matrix with $m = n$
- Identity matrix
  - a square matrix with ones on the main diagonal and zeros everywhere else
  - usually we use a generic notation I (or I_n) to denote an identity matrix
  - $AI = A$
  - $AA^{-1} = I$
- Symmetric matrix
  - a square matrix A is symmetric if $A = A' = A^T$
- Vectorize matrix
  - a matrix A can be reconfigure into a vector: $vec(A_{n \times k}) = nk \times 1$ column vector
  - $\text{vec}\begin{bmatrix} 5 & 2 \\ 9 & 6 \end{bmatrix} = [5, 9, 2, 6]'$

Matrix Inverse and Transpose

- Matrix inverse

$$A^{-1} = \begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$$

A matrix which has an inverse is called to be nonsingular

  - invertible matrix
  - property: $(AB)^{-1} = B^{-1}A^{-1}$

- Transpose Operator
  - turns columns into rows (and vice versa)
  - suppose x is a column vector
  - $x' = x^T = [x_1, x_2, ..., x_n]$: row vector
  - $x'x = \sum_{i=1}^{n} x_i^2$: a scalar
  - $A' = (a_{ji})_{n \times m}$ [recall that $A = (a_{ij})_{m \times n}$]

Some Matrices Operations

- Consider a column vector x (and y) of dimension n and a column vector ι with each element being 1 (same dimension n)
- $\sum_{i=1}^{n} x_i = \iota' x$
- $\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i = a \iota' x$ for a constant a
- If $a = 1/n$, $a \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i / n = \bar{x}$
- $\sum_{i=1}^{n} x_i^2 = x' x$
- $\sum_{i=1}^{n} x_i y_i = x' y = y' x$

Inverse of Partition Matrix

- Inverse of partition matrix

  Let
  $$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$
  If A_{11} and A_{22} are nonsingular, then
  $$\begin{aligned} A^{-1} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} B_{11} & -B_{11}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}B_{11} & A_{22}^{-1} + A_{22}^{-1}A_{21}B_{11}A_{12}A_{22}^{-1} \end{bmatrix}, \end{aligned}$$
  where $B_{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}$.

  - Special case (block diagonal):
  $$\begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & A_{22}^{-1} \end{bmatrix}.$$

Determinant & Trace

- Determinant

$$|A| = \begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad - bc.$$

Properties:

  - $|AB| = |A|\,|B|$
  - $\begin{vmatrix} A_{11} & 0 \\ 0 & A_{22} \end{vmatrix} = |A_{11}|\,|A_{22}|$

- Trace

Trace is the summation of the diagonal elements of a square matrix.

$$\mathrm{tr}\,(A) = \sum_i a_{ii}.$$

  - Important property of the trace: $\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$

Kronecker product

- Kronecker product

  If A is of order $m \times n$ and B is of order $p \times q$, then $C = A \otimes B$ is of order $mp \times nq$. We call matrix C as the Kronecker product of A and B.

$$C = A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \dots & \dots & \dots & \dots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}.$$

  Applications:
  - SUR
  - Panel data

Idempotent Matrix

- Idempotent matrix
  Multiplying matrix A by itself simply reproduces matrix $A \rightsquigarrow AA = A$

Example

Let ι be the $n \times 1$ column vector $[1, 1, ..., 1, 1]'$. Define $P \equiv \iota \left(\iota' \iota\right)^{-1} \iota'$ and $M \equiv I_n - P$.

1. Both P and M are symmetric and idempotent matrices.
2. Furthermore we have $PM = 0$.
3. Premultiplying matrix P will produce the average matrix.
4. Premultiplying matrix M will produce deviation (from mean) matrix.

P Matrix

Example

$\iota'\iota = n$ so $(\iota'\iota)^{-1} = 1/n$. The P matrix looks like this:

$$P = \begin{bmatrix} 1/n & .. & .. & 1/n \\ .. & .. & .. & .. \\ 1/n & .. & .. & 1/n \end{bmatrix}$$

Idempotent Matrix

Example

Projection matrix and residual making matrix in linear regression model

$$
\begin{aligned}
Y &= X\beta + \varepsilon \\
PY &= PX\beta + P\varepsilon = \hat{Y} \\
MY &= MX\beta + M\varepsilon = e \\
P &= X(X'X)^{-1}X' \\
M &= I - P
\end{aligned}
$$

Rank of a Matrix

- Rank of a matrix
  We define the maximum number of linearly independent columns (or rows) in the matrix as the rank of a matrix. Note that # of linearly independent columns must be equal to # linearly independent rows. Rank of a $m \times n$ matrix should satisfy rank $\leq \min(m, n)$. If the rank of a matrix is m, the matrix is said to have full row rank. If the rank of a matrix is n, the matrix is said to have full column rank.
  - Example:

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 1 \\ 2 & 2 & 4 & 5 \end{bmatrix}$$

  - It is obvious that $\text{rank}(A) \leq 3$ or $\rho(A) \leq 3$
  - It turns out that $\text{rank}(A) = 2$ or $\rho(A) = 2$

Rank in Linear Regression Model

- In a linear regression model we have the following data matrix:

$$X_{n\times k} = \begin{bmatrix} 1 \text{ (var1)} & var2 & var3 & \dots & vark \\ \dots & \dots & \dots & \dots & \dots \\ 1 \text{ (var1)} & var2 & var3 & \dots & vark \end{bmatrix}$$

- There are total n observations in row
- There are total k explanatory variables in column
- Since $n \gg k$, rank $\leq \min(n, k)$
- Identification requires full column rank, i.e, $\rho(X) = k$

Eigenvalue & Eigenvector

- Eigenvalue and eigenvector
  Consider a square matrix A, if there are a vector c and a scalar λ such that
  $$Ac = \lambda c.$$
  - uninteresting value for c? can write the equation as
    $$(A - \lambda I)c = 0$$
  - characteristic equation
    $$|A - \lambda I| = 0 \quad (\text{why? if not } c = 0)$$
  - a root of above polynomial equation, λ_i is an eigenvalue of A
  - the corresponding c_i is called an eigenvector of A
  - Example:
    $$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \; B = \begin{bmatrix} 1 & -4 \\ 4 & -7 \end{bmatrix}$$

Why Do We Care about Eigenvalues?

- An $n \times n$ matrix A is positive definite if all eigenvalues of A, $\lambda_1, \lambda_2, ..., \lambda_n$ are positive
  - A matrix is negative-definite, negative-semidefinite, or positive-semidefinite if and only if all of its eigenvalues are negative, non-positive, or non-negative, respectively
- The eigenvectors corresponding to different eigenvalues are linearly independent. So if an $n \times n$ matrix A has n nonzero eigenvalues, it is of full rank
- The trace of a matrix is the sum of the eigenvectors
- The determinant of a matrix is the product of the eigenvectors
- The eigenvectors and eigenvalues of the covariance matrix of a data set are also used in principal component analysis (similar to factor analysis)

Diagonalization

- Diagonalization

  Collecting all n solutions produces the following.

$$A \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} = \begin{bmatrix} \lambda_1 c_1 & \lambda_2 c_2 & \dots & \lambda_n c_n \end{bmatrix}$$
$$= \begin{bmatrix} c_1 & c_2 & \dots & c_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix},$$

  which could be written as

$$AC = C\Lambda.$$

  If C is nonsingular, we will have $C^{-1}AC = \Lambda$ and say that the matrix of eigenvectors serves to diagonalize the A matrix.

  - For a symmetric matrix A, it turns out that eigenvectors are orthogonal and we can make them normalized. (orthonormal)

Properties of Eigenvalues

- Properties of Eigenvalues
  - Eigenvalues of a symmetric matrix are all real.
  - Eigenvalues of a nonsymmetric matrix may be real or complex.
  - The rank of A is equal to the number of nonzero eigenvalues.
  - $|A| = \left|C^{-1}\Lambda C\right| = \prod_{i=1}^{n} \lambda_i$
    - non-zero eigenroot and non-singular matrix
  - $\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$
  - Every eigenroot of an idempotent matrix is either 0 or 1.
    - note that $A^h c = \lambda^h c$
    - for instance, $Ac = \lambda c = A^2 c = A\lambda c = \lambda^2 c$

Quadratic Forms

- Quadratic forms
  Consider a $n \times n$ symmetric matrix A and a $n \times 1$ vector $\mathbf{x}$, the scalar

$$q = \mathbf{x}'A\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

  is called the quadratic form. If $q \geq 0\,(q > 0)$ for any $x \neq 0$, then A is said to be positive semi-definite (positive definite). If $q \leq 0\,(q < 0)$ for any $x \neq 0$, then A is said to be negative semi-definite (negative definite). The idea above is quite similar to the scalar case. If matrix difference $A - B$ is p.d. (p.s.d.), we say that $A > B$ $(A \geq B)$.
- $q_A = x_1^2 + x_2^2$, $q_B = (x_1 + x_2)^2$, $q_C = x_1^2 - x_2^2$

Matrix Differentiation

- Matrix differentiation
  - (Case I) scalar f and $n \times 1$ vector $\mathbf{x}$

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \dots \\ f_n \end{bmatrix}$$

$$\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}'} = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & f_{2n} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nn} \end{bmatrix}, \text{ where } f_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

  - If $f = a'\mathbf{x} = \mathbf{x}'a$, $\partial(a'\mathbf{x})/\partial \mathbf{x} = \partial(\mathbf{x}'a)/\partial \mathbf{x} = a$.

Matrix Differentiation

Example

If β and a are both $k \times 1$ vectors then, $\frac{\partial \beta' a}{\partial \beta} = a$

Proof.

$$
\begin{aligned}
\frac{\partial \beta' a}{\partial \beta} &= \frac{\partial}{\partial \beta}(\beta_1 a_1 + \beta_2 a_2 + \ldots + \beta_k a_k) \\
&= \begin{bmatrix} \frac{\partial}{\partial \beta_1}(\beta_1 a_1 + \beta_2 a_2 + \ldots + \beta_k a_k) \\ \frac{\partial}{\partial \beta_2}(\beta_1 a_1 + \beta_2 a_2 + \ldots + \beta_k a_k) \\ \ldots \\ \frac{\partial}{\partial \beta_k}(\beta_1 a_1 + \beta_2 a_2 + \ldots + \beta_k a_k) \end{bmatrix} \\
&= a
\end{aligned}
$$

□

Matrix Differentiation

- Matrix differentiation
  - (Case II) $m \times 1$ vector f and $n \times 1$ vector $\mathbf{x}$
    Gradient is

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1m} \\ f_{21} & f_{22} & \dots & f_{2m} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nm} \end{bmatrix}_{n \times m}, \text{ where } f_{ij} = \frac{\partial f_j}{\partial x_i}$$

    - If $f = A\mathbf{x}$, $\partial(A\mathbf{x})/\partial\mathbf{x} \equiv \partial(\mathbf{x}'A')/\partial\mathbf{x} = A'$.
  - (Case III) scalar f and $n \times k$ matrix A
    Gradient is

$$\frac{\partial f}{\partial A} = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1k} \\ f_{21} & f_{22} & \dots & f_{2k} \\ \dots & \dots & \dots & \dots \\ f_{n1} & f_{n2} & \dots & f_{nk} \end{bmatrix}, \text{ where } f_{ij} = \frac{\partial f}{\partial a_{ij}}$$

Matrix Differentiation

Example

If β be a $k \times 1$ vector and A be a $n \times k$ matrix then $\frac{\partial A\beta}{\partial \beta'} = A$.

Proof.

$$\frac{\partial A\beta}{\partial \beta'} = \frac{\partial}{\partial \beta'} \begin{bmatrix} a_{11}\beta_1 + a_{12}\beta_2 + \ldots + a_{1k}\beta_k \\ a_{21}\beta_1 + a_{22}\beta_2 + \ldots + a_{2k}\beta_k \\ \ldots \\ a_{n1}\beta_1 + a_{n2}\beta_2 + \ldots + a_{nk}\beta_k \end{bmatrix}$$

$$= \begin{bmatrix} [\ \frac{\partial}{\partial \beta_1} \quad \cdots \quad \frac{\partial}{\partial \beta_k}\ ] a_{11}\beta_1 + a_{12}\beta_2 + \ldots + a_{1k}\beta_k \\ [\ \frac{\partial}{\partial \beta_1} \quad \cdots \quad \frac{\partial}{\partial \beta_k}\ ] a_{21}\beta_1 + a_{22}\beta_2 + \ldots + a_{2k}\beta_k \\ \ldots \\ [\ \frac{\partial}{\partial \beta_1} \quad \cdots \quad \frac{\partial}{\partial \beta_k}\ ] a_{n1}\beta_1 + a_{n2}\beta_2 + \ldots + a_{nk}\beta_k \end{bmatrix} = A$$

$\square$

Matrix Differentiation

- Matrix differentiation

  Properties:

  - $\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} = A'$
  - $\frac{\partial \mathbf{x}' A\mathbf{x}}{\partial \mathbf{x}} = (A + A')\,\mathbf{x}$
  - $\frac{\partial \mathbf{x}' A\mathbf{x}}{\partial \mathbf{x}} = 2A\mathbf{x}$ if A is symmetric.

Matrix Differentiation

Example

Let β be a 2×1 vector and A be a 2×2 symmetric matrix then $\frac{\partial \beta' A \beta}{\partial \beta} = 2A\beta$

Proof.

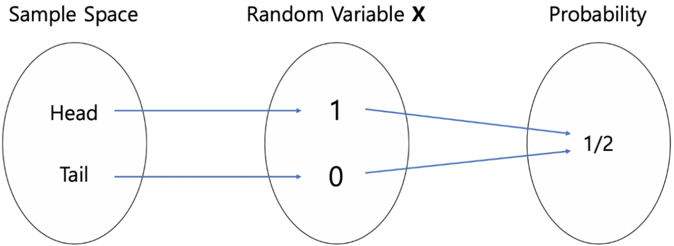
$$\begin{aligned}
\frac{\partial \beta' A \beta}{\partial \beta} &= \frac{\partial}{\partial \beta}(\beta_1^2 a_{11} + 2a_{12}\beta_1\beta_2 + \beta_2^2 a_{22}) \\
&= \begin{bmatrix} \frac{\partial}{\partial \beta_1}(\beta_1^2 a_{11} + 2a_{12}\beta_1\beta_2 + \beta_2^2 a_{22}) \\ \frac{\partial}{\partial \beta_2}(\beta_1^2 a_{11} + 2a_{12}\beta_1\beta_2 + \beta_2^2 a_{22}) \end{bmatrix} \\
&= \begin{bmatrix} 2\beta_1 a_{11} + 2a_{12}\beta_2 \\ 2\beta_1 a_{12} + 2a_{22}\beta_2 \end{bmatrix} = 2A\beta
\end{aligned}$$

□

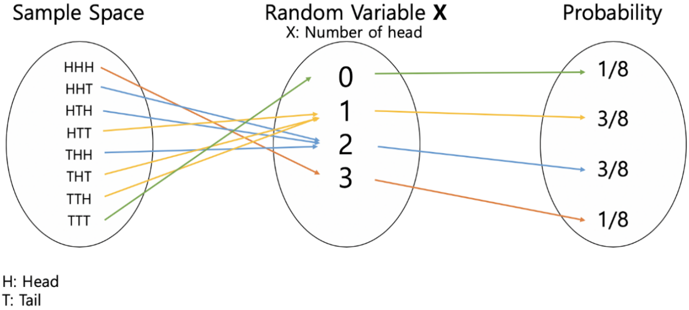
Some Concepts

- Population vs. sample
- Random variable
  R.V. is a variable which we assign a set of possible values and associated probabilities.
  - Example: $\Pr(\text{Head}) = \Pr(X = 1) = 1/2$.
- Probability density function: $f(x)$
  $f(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ and $\int_a^b f(x)\mathrm{d}x = \Pr[a < x < b]$.
- Cumulative density function: $F(x)$
  $F(x) = \Pr[X \leq x] \geq 0$ for all x
  $\int_{-\infty}^{x} f(x)\mathrm{d}x = F(x)$
- Expectation and variance
  $\mathrm{E}[X] = \int xf(x)\mathrm{d}x = \mu$
  $\mathrm{var}[X] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2 = \mathrm{E}[X^2] - \mu^2$

Random Variable I



Random Variable II



Some Concepts

- Chebyshev inequality

$$\Pr\left[|X - \mathrm{E}[X]| > t\right] \leq \frac{\mathrm{var}[X]}{t}, \text{ for any } t > 0$$

  - other inequalities in asymptotics later on

- Cauchy-Schwartz inequality

$$\mathrm{E}[XY]^2 \leq \mathrm{E}[X^2]\mathrm{E}[Y^2]$$

  - Let X and Y be zero mean variables
  - correlation coefficient?
  - analogy principle; sample counterpart?

$$\sum_{i=1}^{n}[X_i Y_i]^2 \leq \sum_{i=1}^{n}[X_i^2]\sum_{i=1}^{n}[Y_i^2]$$

Delta Method

- Delta method

  Given $\mathrm{E}[X] = \mu$. What are the expectation and variance of $g(X)$? By linear Taylor expansion, we have

$$g(X) \simeq g(\mu) + g'(\mu)(X - \mu).$$

  Therefore, one could approximate $\mathrm{E}[g(X)]$ and $\mathrm{var}[g(X)]$ by the following:

$$\begin{aligned} \mathrm{E}[g(X)] &\simeq g(\mu) \\ \mathrm{var}[g(X)] &\simeq \left[g'(\mu)\right]^2 \mathrm{var}[X] \end{aligned}$$

Uncorrelatedness and Independence

- Uncorrelatedness and Independence
  - (Stochastic) Independence implies uncorrelatedness but not vice versa.
  - $\text{cov}(X, Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y) = ...$

Law of Iterative Expectation & Decomposition of Variance

- Law of Iterative Expectation (LIE or Double Expecations)
  $\mathrm{E}[Y] = \mathrm{E}_X[\mathrm{E}[Y|X]]$
  - Proof:
    $\int\int yf(x,y)\mathrm{d}x\mathrm{d}y = \int\int yf(y|x)f(x)\mathrm{d}x\mathrm{d}y = \int \mathrm{E}[y|x]f(x)\mathrm{d}x$
  - if $\mathrm{E}[Y|X] = 0 \Rightarrow \mathrm{E}[Y] = 0$, $\mathrm{E}[XY] = 0$, and $\mathrm{cov}[X,Y] = 0$
- Decomposition of Variance
  $\mathrm{var}[y] = \mathrm{var}_x[\mathrm{E}[y|x]] + \mathrm{E}_x[\mathrm{var}[y|x]]$

Normal & Chi-square Distributions

- Normal distribution
  If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, the normal density is

$$f\left(x|\mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

- Special case: standard normal – $X \sim \mathcal{N}(0, 1)$.
- Multivariate normal – $X \sim \mathcal{N}(\mu, \Sigma)$.

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right],$$

where Σ is the variance-covariance matrix.

Chi-square Distribution

- Chi-square distribution
  If $Z \sim \mathcal{N}(0,1)$, then $X = Z^2 \sim \mathcal{X}^2(1)$.
- Sum of $\mathcal{X}^2$ distributions: If $x_1, ..., x_n$ are n independent $\mathcal{X}^2(1)$, then $\sum_{i=1}^{n} X_i \sim \mathcal{X}^2(n)$.
- Furthermore, we have $\mathrm{E}\left[\mathcal{X}^2(n)\right] = n$ and $\mathrm{var}\left[\mathcal{X}^2(n)\right] = 2n$.
- If $W \sim \mathcal{N}(\mathbf{0},\Sigma)$, what is the distribution of $W'\Sigma^{-1}W$?
- If W_i's are uncorrelated and have the distribution $\mathcal{N}\left(0,\sigma_i^2\right)$, what is the distribution of $W'\Sigma^{-1}W$?
  - Wald-type statistic when we talk about trinity of test

F & t Distributions

- F distribution
  If x_1 and x_2 are two independent Chi-squared random variables with degrees of freedom n_1 and n_2, respectively, then the ratio

$$F(n_1, n_2) = \frac{x_1 / n_1}{x_2 / n_2}$$

  is the F distribution with n_1 and n_2 degrees of freedom.
  - convergence of $n_1 F$ as $n_2 \to \infty$?
- t distribution
  If $Z \sim \mathcal{N}(0,1)$, and $X = \mathcal{X}^2(n)$ is independent of Z, then t distribution with n degrees of freedom is defined below.

$$t(n) = \frac{Z}{\sqrt{X/n}}.$$

- Other distributions: Gamma, Beta, Logistic, Bernoulli, Binomial, Poisson,..., etc.

Some Univariate Distributions

Example

Consider a random sample X_i drawing from the uniform distribution $\mathcal{U}[0,1]$. Compute $\mathrm{E}[X_i]$ and $\mathrm{var}[X_i]$.

Example

Consider a random sample Y_i drawing from the exponential distribution $\mathcal{E}xp[1]$. Compute $\mathrm{E}[Y_i]$ and $\mathrm{var}[Y_i]$. [Note: The density of Y_i is given by $f(y_i) = \exp(-y_i)$.]

Bivariate Distributions

- Bivariate distributions

  Joint density of x and y: $f(x, y)$. The joint probability is defined as

  $$\Pr[a < x \leq b, c < y \leq d] = \int_a^b \int_c^d f(x, y)\, dy dx.$$

  Marginal density of x is

  $$f_x(x) = \int_{-\infty}^{\infty} f(x, y)\, dy.$$

  - If x and y are independent, then $f(x, y) = f_x(x) f_y(y)$.

Transformation of Random Variables

- Transformation of random variables
  If x is a continuous variable with density $f_x(x)$, and $y = g(x)$ is a continuous *monotonic* function of x, then the density of y could be obtained by

$$f_y(y) = f_x\left[g^{-1}(y)\right]\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|.$$

  - In many application, the function g may be monotone over certain intervals. If this is the case, we have to handle it interval by interval. For instance, $Y = X^2$.

Transformation of Random Variables

Example

Let $x \sim \mathcal{N}(\mu, \sigma^2)$, we are interested in the density of $y = (x - \mu)/\sigma$. Clearly, $g^{-1}(y) = x = \sigma y + \mu$. Jacobian term is $\left| \mathrm{d}g^{-1}(y)/\mathrm{d}y \right| = |\sigma|$. Hence,

$$f_y(y) = f_x[\sigma y + \mu]\,|\sigma| = \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\sigma y)^2}{2\sigma^2} \right] \right] |\sigma|$$

Sampling Distribution

- Sampling distribution

  A sample of n observations, $x_1, ..., x_n$, is a random sample if n observations are drawn *independently* from the same (*identical*) density, $f(x)$. Note that the abbreviation: *i.i.d. (independently identically distributed)*

Example

$x_1, ..., x_n$, i.i.d. from $f(x) = \theta \exp[-\theta x]$ with $0 < x < \infty$. Please find out the sampling distribution of the sample minimum, $x_{(1)}$.

Estimator

- Estimator
  In practice, we use a function of data to estimate the population parameters. The function of data is called estimator.

Example

One may use sample mean $\bar{x}$ (or $x_{(1)}$) to estimate population parameter μ. Which's better?

Finite Sample Properties of Estimator

- Finite sample properties of estimator
  1. Unbiased estimator: $\mathrm{E}[\hat{\theta}] = \theta$.
  2. Efficient unbiased estimator: For unbiased estimator $\hat{\theta}_1$ and $\hat{\theta}_2$, if $\mathrm{var}[\hat{\theta}_1] < \mathrm{var}[\hat{\theta}_2]$, we say that unbiased estimator $\hat{\theta}_1$ is more efficient than another unbiased estimator $\hat{\theta}_2$. For vector case, we need $\mathrm{var}[\hat{\theta}_1] - \mathrm{var}[\hat{\theta}_2]$ to be n.s.d.
  3. Mean squared error (MSE):

$$\begin{aligned}\mathrm{MSE}[\hat{\theta}] &= \mathrm{E}[(\hat{\theta} - \theta)^2] \\ &= \mathrm{var}[\hat{\theta}] + [\mathrm{E}[\hat{\theta}] - \theta]^2 (\text{why?}) \\ &= \mathrm{var}[\hat{\theta}] + \mathrm{Bias}[\hat{\theta}]^2.\end{aligned}$$

MLE

- Maximum likelihood estimation [MLE]
  $x_1, ..., x_n$, is an i.i.d. random sample drawn from $f(x_i, \theta)$. The likelihood function (joint density) is

$$f(x_1, x_2, ...x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \mathcal{L}(\theta | x_1, x_2, ...x_n).$$

Example

The likelihood function of exponential distribution is

$$\mathcal{L}(\theta | x_1, x_2, ...x_n) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}.$$

CRLB

- Cramér-Rao Lower Bound [CRLB]
  Under some regularity conditions, the variance of an unbiased estimator of parameter θ will be at least as large as the inverse of information matrix $(\mathcal{I}(\theta))$.

$$[\mathcal{I}(\theta)]^{-1} = \left[-\mathsf{E}\left[\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2} \right] \right]^{-1} = \left[\mathsf{E}\left[\left(\frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta} \right)^2 \right] \right]^{-1}.$$

Example of MLE and CRLB

- Consider random sampling a normal distribution and derive the variance bound.
- Take two unbiased estimator $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n - 1$. (why?)
- Do they achieve CRLB?
- Log-Likelihood function:

$$\ln \mathcal{L}(\mu, \sigma^2) = -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Example of MLE and CRLB

- Example of MLE and CRLB
  - FOCs:

$$\frac{\partial \ln \mathcal{L}(\theta)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu),$$

$$\frac{\partial \ln \mathcal{L}(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \mu^2} = -\frac{n}{\sigma^2},$$

$$\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)$$

Example of MLE and CRLB

- Example of MLE and CRLB - Continued
  - Information matrix is
$$\mathcal{I}(\theta) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$
  - Now we know $\mathcal{I}(\theta)^{-1}$ is the variance bound of unbiased estimators of μ and σ^2.
$$\mathcal{I}(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}.$$
  - Variance matrix of $\hat{\mu}$ and $\hat{\sigma}^2$ is
$$\text{var}[\hat{\theta}] = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{bmatrix}.$$

Example of MLE and CRLB

- Example of MLE and CRLB - Continued
  - Therefore,

$$\text{var}[\hat{\theta}] - \mathcal{I}(\theta)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{2\sigma^4}{(n-1)n} \end{bmatrix}.$$

Conclusion is that $\hat{\mu}$ attains CRLB but not $\hat{\sigma}^2$.